PLANT SCIENCE TODAY
Vol x(x): xx–xx
https://doi.org/10.14719/pst.8036

eISSN 2348-1900

RESEARCH ARTICLE

# Evaluating the effectiveness of principal component regression vs multiple linear regression for black gram cultivation in Tamil Nadu

**Vasanthi R[1*], Karthick V[2], Nirmala Devi M[1], Chellamuthu R[1] & Hema Bharathi C[1]**

[1]Department of Physical Sciences & IT, Agricultural Engineering College and Research Institute, Coimbatore 641 003, Tamil Nadu, India
[2]Centre for Agricultural and Rural Development Studies, Department of Agricultural Economics, Tamil Nadu Agricultural University, Coimbatore 641 003, Tamil Nadu, India

*Correspondence email - vasanthi@tnua.ac.in

Received: 03 March 2025; Accepted: 26 March 2025; Available online: Version 1.0: 25 July 2025

**Cite this article:** Vasanthi R, Karthick V, Nirmala DM, Chellamuthu R , Hema BC. Evaluating the effectiveness of principal component regression vs multiple linear regression for black gram cultivation in Tamil Nadu. Plant Science Today (Early Access). https:/doi.org/10.14719/pst.8036

## Abstract

This study examines the comparison between Multiple Linear Regression and Principal Component Regression for black gram cultivation in Tamil Nadu. This research addresses the problem of accurately modeling and predicting the factors that influence the yield and productivity of black gram cultivation in Tamil Nadu. The challenge lies in identifying which statistical technique, Principal Component Regression (PCR) or Multiple Linear Regression (MLR), is more effective in capturing the complex relationships between various environmental, economic and agricultural variables that affect black gram production. Secondary data were collected from 1999 to 2022 (23 years). Considering yield as a dependent variable and the independent variables are Seed, Fertilizer, Manure, Human labour and Animal labour. According to Multiple Linear Regression, the regression coefficients of fertilizer and human labour significantly influence the yield. The coefficients of Fertilizer and Human labour are found to be 0.049 and 0.07 respectively. According to Principal Component Analysis, the 2 principal components are chosen because the eigenvalue is more than 1.0. These 2 components, PC1 and PC2, cover 36 % and 34 % respectively. The loadings revealed that fertilizer, manure and animal labour significantly contributed to PC1 and Seed and human labour contributed significantly in PC2. The Multiple Linear Regression and Principal Component regression are compared using R square, Adjusted R Square, Root Mean Square, Mean Absolute Error and Mean Absolute Percentage Error. The adjusted R square reveals that Principal Component Regression is better than Multiple Linear Regression. The lowest value of Root Mean Square Error, Mean Absolute Error and Mean Absolute Percentage Error shows the best model among the 2 models. The error is lower for Principal Component Regression compared to Regression. PC1 captures the relationship between fertilizer, manure and animal labor, representing an "input utilization efficiency" dimension. PC2 reflects the trade-off between human labor and seed usage, defining a spectrum between "labor-intensive" and "seed-reliant" farming strategies.

**Keywords:** black gram; green gram; multiple linear regression; principal component

## Introduction

India is the largest producer and consumer of pulses in the world, where pulses play a crucial role in agriculture. The country's total pulse production stands at 23.95 million tons. Black gram (*Vigna mungo*), commonly known as urad dal, is a crucial pulse crop in India, both in terms of domestic consumption and export potential. India is the largest producer and consumer of black gram, contributing over 70 % of global production. The crop plays a vital role in crop rotation, improving soil fertility and stabilizing farmer income.

Black gram, commonly known as urad dal, is a significant pulse crop in India, contributing substantially to the country's pulse production. In the 2022-23 agricultural year, India's black gram (urad) production was estimated at 2.631 million tonnes. For the 2023-24 period, production is expected to decrease to 2.055 million tonnes. This decline is attributed to a reduction in cultivation area and adverse weather conditions that have affected crop yields. Notably, Madhya Pradesh led production with 473,000 tonnes, followed by Uttar Pradesh at 442,000 tonnes and Andhra Pradesh contributing 333,000 tonnes. Despite the challenges, India remains the largest producer and consumer of black gram, with a significant portion of the crop dedicated to domestic demand for traditional dishes like dal makhani and idli batter.

Over the past years, Tamil Nadu has demonstrated consistent growth in black gram (urad dal) production, driven by favourable climatic conditions, government incentives and the adoption of high-yielding seed varieties. In 2022-23, the state produced approximately 5.7 lakh metric tonnes, reflecting a slight increase from the 5.4 lakh metric tonnes recorded in 2021-22. The districts of Thanjavur, Nagapattinam and Cuddalore remained key contributors to the output, benefiting from well-irrigated delta regions. Additionally, the

Tamil Nadu government's promotion of pulse cultivation under schemes like PMKSY (Pradhan Mantri Krishi Sinchayee Yojana) and subsidies for drip irrigation significantly boosted productivity and farmer participation.

The problem of food and nutritional security, particularly in Tamil Nadu, is multifaceted and directly linked to agricultural production, especially that of pulses such as black gram (urad dal). Black gram, scientifically known as *V. mungo*, belongs to the family Fabaceae (Leguminosae) and is a significant pulse crop in India. The plant exhibits a racemose type of inflorescence, meaning that its flowers are arranged in clusters on long stems (1). The fruit produced by black gram is a legume pod, typically 5 -7 cm long, containing 2 - seeds (2). These seeds are small, oval-shaped and black, although the color can range from dark brown to light gray in some varieties. India remains the largest producer and consumer of black gram globally and the crop plays a critical role in domestic nutrition, contributing essential protein to the diets of millions of Indians (3).

In Tamil Nadu, although there has been consistent growth in black gram production due to favorable climatic conditions, government incentives and improved farming practices, the state's production remains vulnerable to several factors (1). A notable challenge is the fluctuation in yield due to adverse weather conditions such as droughts, unseasonal rains, or floods, which directly impact the quantity and quality of pulses produced (4). For example, in the 2023 - 24 period, there was a decrease in black gram production, primarily due to reduced cultivation areas and climatic stress (5). This reduction in output threatens the stability of supply and may lead to higher prices, affecting the affordability and accessibility of black gram for households.

Black gram provides a rich source of protein (25.2 g), dietary fiber (18.4 g) and essential minerals like iron (7.57 mg), calcium (138 mg) and magnesium (192 mg). These nutrients are essential for various bodily functions, including muscle repair, immune system support, bone health and red blood cell production. Additionally, black gram is high in folate (276 µg), which is crucial for cell division and the prevention of neural tube defects during pregnancy (6).

The objective of this study is to compare Principal Component Analysis (PCA) and Multiple Linear Regression (MLR) in evaluating key factors influencing black gram cultivation in Tamil Nadu. It aims to identify significant factors that affect/influencing yield, PCA will effectively reduce the dimensionality of the dataset and identify the most important factors influencing black gram yield in Tamil Nadu (7). At the same time, MLR will provide a robust predictive model for assessing the relationships between these factors and yield. It is hypothesized that PCA will outperform MLR in terms of model accuracy and computational efficiency, thereby offering a more effective approach for optimizing black gram production (8).

## Materials and Methods

Secondary data covering 23 years (1999 - 2022) was collected from the Season and Crop Report, Tamil Nadu. Yield is considered a dependent variable, while the independent variables include seed, fertilizer, manure, human labor and animal labor. In this study, we employed MLR to examine the relationship between the dependent variable (Y) and a set of independent variables. The MLR model is given by the equation:

$$Y = a0+b1x1+b2x2+b3x3+b4x4+b5x5+e$$

where '$a_0$' represents the intercept, b1, b2, b3, b4 & b5 are the regression coefficients for the independent variables and "e" is the error term. The model was used to assess the impact of independent variables on the dependent variable (9,10).

PCA is a widely used dimensionality reduction technique in statistics and machine learning. It transforms a dataset with correlated variables into a set of uncorrelated variables, known as principal components (PCs) (11). The goal is to capture the maximum variance in fewer dimensions, simplifying analysis while preserving essential patterns in the data. PCA helps in simplifying large datasets by reducing the number of variables while retaining important information. Since PCA generates uncorrelated principal components, it eliminates redundancy in correlated variables. Reducing the number of variables enhances the computational efficiency of machine learning models and statistical analyses (12).

The working procedure is given below:

**1.Standardization (Pre-processing the data)**

- PCA is sensitive to the scale of variables, so data is standardized (normalized) to have a mean of zero and a standard deviation of one (Z-score normalization) (11).
- Standardization ensures that all variables contribute equally to the analysis (11).

**2.Compute the covariance matrix**

- The covariance matrix captures the relationships between all variables. If two variables are highly correlated, PCA will likely combine them into a single principal component (11).

**3.Perform eigen decomposition**

- Eigenvalues represent the amount of variance explained by each principal component.
- Eigenvectors define the directions (axes) of the new feature space (12).

**4.Select the number of principal components**

- **Explained variance ratio**: A cumulative threshold (e.g., 85 % - 95 %) is used to decide how many principal components to retain.
- **Scree plot**: A visual graph helps determine the "elbow point," where adding more components has little effect on variance explained (11).

**5.Transform the original data**

- The original data is projected onto the principal components to create a new set of uncorrelated variables.
- Each principal component (PC) is a linear combination of the original variables. The first PC (PC1) captures the most variance, followed by PC2, PC3, etc. The weight (loading) of each variable in a principal component determines its

importance (11,12).

Finally, results are analyzed using biplots, loadings and explained variance and the transformed data is used for further statistical modeling or machine learning applications (11,12).

## Results and Discussion

### Multiple linear regression

Multiple linear regression and principal Ccmponent analysis were performed for the collected data (13). Considering yield as a dependent variable and the independent variables are Seed, Fertilizer, Manure, Human labour and Animal labour.

Table 1, shows the R square and Adjusted R square values are 0.65 and 0.54 respectively. The regression model explains 65 % of the variability in yield (14).

In the ANOVA Table 2, the p-value (0.002) of the regression is less than 0.05, which confirms the significance of the regression of the given data. To study the impact of input variables MLR is performed in SPSS and the results of the MLR table are given in Table 3.

It could be observed from Table 3 among 5 independent variables, the regression coefficients of fertilizer and human labour are significantly influencing the yield, which corroborates findings in previous studies on crop yield predictions (15,16). The coefficients of Fertilizer and human labor are 0.049 and 0.07.

The residuals of regression are plotted as a histogram and a Normal P-P plot in Fig. 1 and Fig. 2, used to verify the validation of the regression model. It revealed that the residual follows a normal distribution with zero mean and variance.

### Principal component regression

Principal component regression is used to overcome the multicollinearity of explanatory variables and reduce the dimensionality of variables.

Principal component analysis is performed for independent variables. PCA (17) is carried out in SPSS-22 version. The eigenvalues of principal components are denoted in the scree plot (Fig. 3). The 2 principal components are chosen because the eigenvalue is more than 1.0 (Table 4). These 2 components, PC1 and PC2, cover 36.13 % and 33.72 %, respectively. The loadings of principal components indicate how much each original variable contributes to PC1 and PC2 and are given in the component matrix (Table 5), which is used to calculate principal scores (18). The loadings revealed that fertilizer, manure and animal labor significantly contributed to PC1, while seed and human labor contributed significantly to PC2. Additionally, animal labor significantly contributed to PC1 and Seed and human labor contributed significantly to PC2 (19).

**Table 1.** R-square & adjusted R-square

| Model | R | R square | Adjusted R square | Std. error of the estimate |
|---|---|---|---|---|
| 1 | .804[a] | .647 | .543 | .89578 |

**Table 2.** ANOVA

| ANOVA[a] | | | | | | |
|---|---|---|---|---|---|---|
| Model | | Sum of squares | df | Mean square | F | Sig. |
| 1 | Regression | 24.984 | 5 | 4.997 | 6.227 | .002[b] |
| | Residual | 13.641 | 17 | .802 | | |
| | Total | 38.625 | 22 | | | |

In principal component regression, principal scores are utilized as the independent variable, keeping yield as a dependent variable. From the Table 6 shows R square and Adjusted R square of this regression are 0.635 and 0.577, respectively, which aligns with previous research on crop yield predictions using PCA (20). The ANOVA Table 7 confirms that regression is significant for the given data. The p-value of the t-test (Table 8) shows that the regression coefficient of PC1 significantly influences the yield. The regression coefficients of PC1 and PC2 scores are 0.058 and -0.018, respectively. It is noticeable that the PC2 is not significant.

### Multiple linear regression Vs principal component regression

The regression and principal component regression are compared using R square, Adjusted R Square, Root Mean Square, Mean Absolute Error and Mean Absolute Percentage Error. It could be observed from Table 9 that the large R Square and Adjusted R square explain the maximum variability of the dependent variable. The adjusted R square reveals that principal component regression is better than multiple linear regression, confirming similar findings in pulse crops (21). The

**Table 3.** MLR results

| Coefficients[a] | | | | | | |
|---|---|---|---|---|---|---|
| Model | | Unstandardized Coefficients | | Standardized Coefficients | T | Sig. |
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | .010 | 3.657 | | .003 | .998 |
| | Seed (Kg/hec) | .070 | .125 | .104 | .559 | .583 |
| | Fertilizer (Kg/hec) | .049 | .020 | .564 | 2.395 | .028 |
| | Manure (Qtl/hec) | .012 | .034 | .061 | .349 | .731 |
| | Human_labour(hrs) | .007 | .002 | .568 | 3.210 | .005 |
| | Animal_labour (hrs) | -.020 | .053 | -.092 | -.374 | .713 |
| a. Dependent variable: Yield | | | | | | |

**Table 4.** Principal component analysis

| Total Variance Explained | | | | | | |
|---|---|---|---|---|---|---|
| Component | Initial Eigen values | | | Extraction Sums of Squared Loadings | | |
| | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 1.806 | 36.128 | 36.128 | 1.806 | 36.128 | 36.128 |
| 2 | 1.686 | 33.720 | 69.848 | 1.686 | 33.720 | 69.848 |
| 3 | .862 | 17.247 | 87.094 | | | |
| 4 | .472 | 9.445 | 96.539 | | | |
| 5 | .173 | 3.461 | 100.000 | | | |

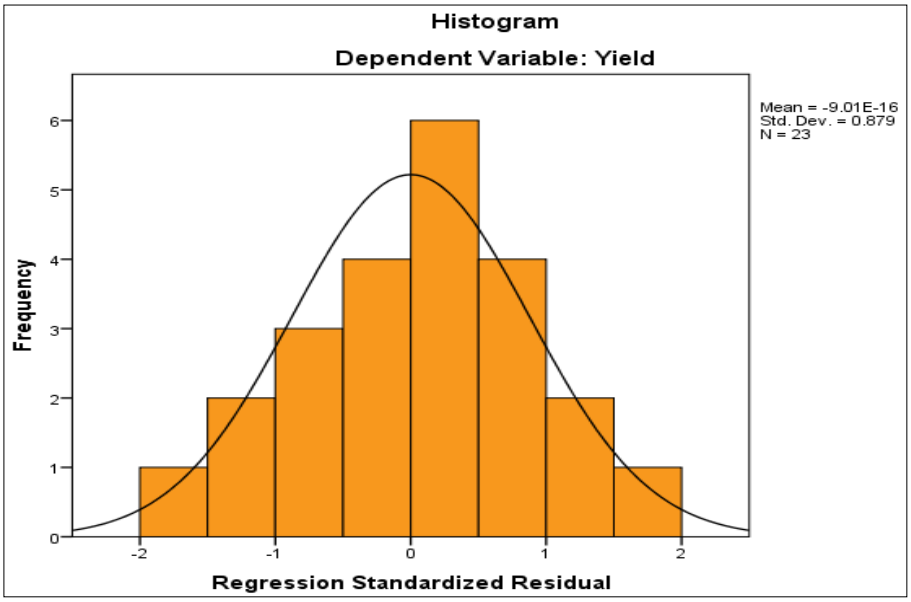


**Fig.1.** Histogram.

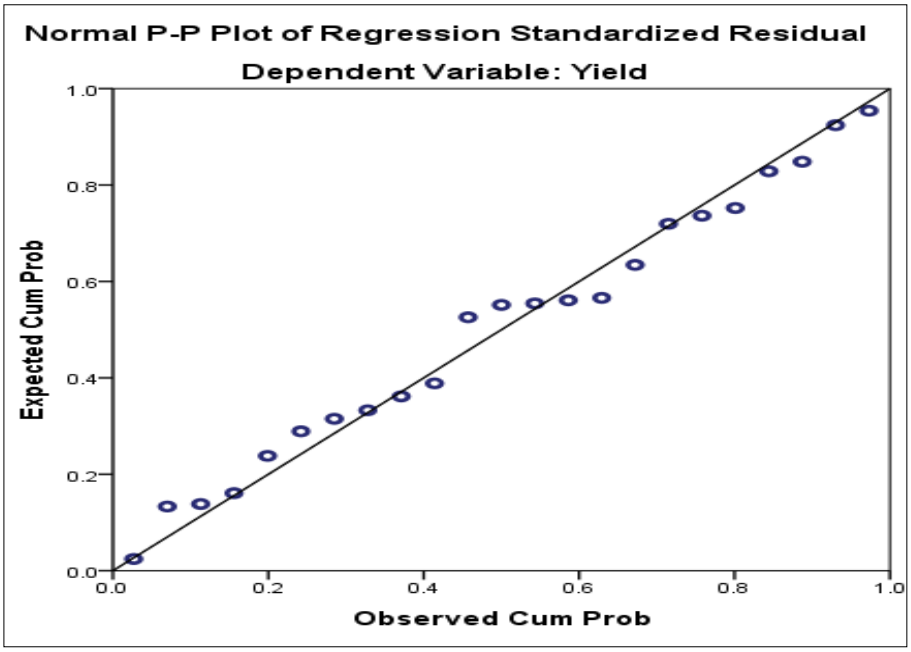


**Fig.2.** Normal P-P plot.

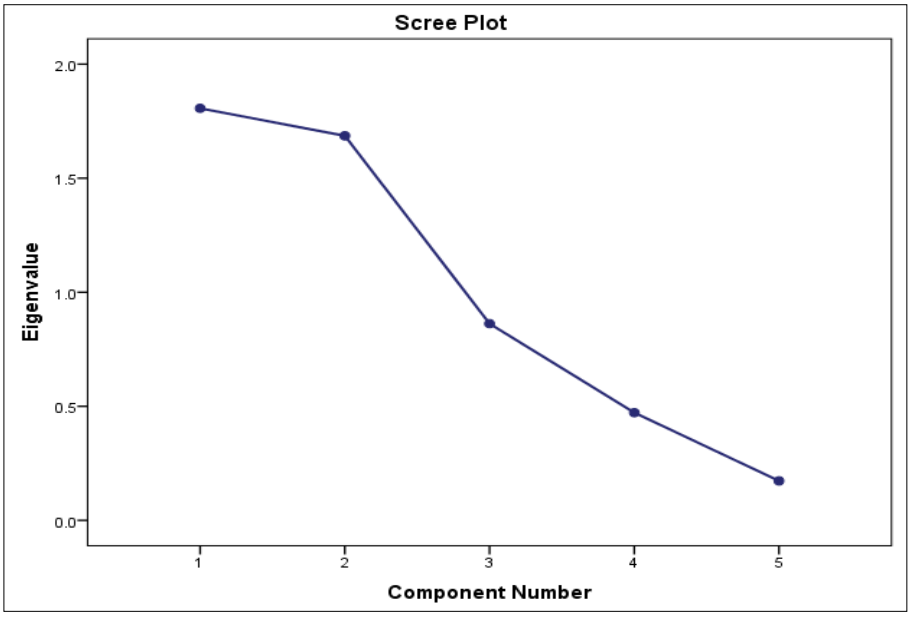


**Fig.3.** Scree plot.

**Table 5.** Component matrix

| Component matrix[a] | | |
|---|---|---|
| | Component | |
| | 1 | 2 |
| Seed | -.303 | -.628 |
| Fertilizer | .746 | -.484 |
| Manure | .663 | .455 |
| Human_labour | .391 | .735 |
| Animal_labour | -.751 | .557 |

**Table 6.** Model summary

| Model summary[b] | | | | |
|---|---|---|---|---|
| Model | R | R square | Adjusted R square | Std. error of the estimate |
| 1 | .797[a] | .635 | .577 | .86176 |

**Table 7.** ANOVA

| ANOVA[a] | | | | | | |
|---|---|---|---|---|---|---|
| Model | | Sum of squares | df | Mean square | F | Sig. |
| 1 | Regression | 24.515 | 3 | 8.172 | 11.004 | .000[b] |
| | Residual | 14.110 | 19 | .743 | | |
| | Total | 38.625 | 22 | | | |

**Table 8.** Principal component regression coefficients

| Coefficients[a] | | | | | | |
|---|---|---|---|---|---|---|
| Model | | Unstandardized coefficients | | Standardized coefficients | t | Sig. |
| | | B | Std. error | Beta | | |
| 1 | (Constant) | 2.343 | .947 | | 2.473 | .023 |
| | P1 | .058 | .012 | 1.458 | 4.849 | .000 |
| | P2 | -.018 | .027 | -.819 | -.651 | .523 |

**Table 9.** Model comparison

| Model | R square | Adjusted R square | RMSE | MAE | MAPE |
|---|---|---|---|---|---|
| Regression | 0.647 | 0.543 | 0.7701 | 0.6219 | 12.51619 |
| Principal component regression | 0.635 | 0.577 | 0.7833 | 0.5925 | 11.9343 |

lowest value of RMSE, MAE and MAPE shows the best model among the 2 models (22). The MAE and MAPE are lower for principal component regression compared to regression.

## Conclusion

According to the results of Multiple linear regression and principal component regression performs well with 2 components principal component regression, it is found that principal component regression performs well with 2 components, PC1 and PC2. The MAE and MAPE are lower for principal component regression compared to multiple linear regression. The 2 components, PC1 and PC2, together capture nearly 70 % of the variance in the independent variables, making them the primary dimensions for analysis. Fertilizer, manure and animal labor (captured in PC1) are the most important contributors to yield. Efforts to optimize these inputs could significantly boost productivity. PC1 represents the interplay between inputs like fertilizer, manure and animal labor, indicating an "input efficiency" dimension and PC2 highlights a balance between human labor and seed usage, suggesting a "labor-focused vs. seed-intensive" farming strategy. Future research could expand beyond PCA and MLR by exploring machine learning models like random forests, support vector machines, or neural networks, which are capable of identifying non-linear relationships and complex interactions between variables that PCA and MLR may not capture.

## Acknowledgements

We would like to express our sincere gratitude to the Tamil Nadu Agricultural University for providing the data from their Season and Crop Report, which was integral to this study. Additionally, we acknowledge their support in conducting the research and ensuring its success.

## Authors' contributions

VR carried out the design of the study with necessary objectives developing the concept for the research, including its goals, structure and approach. KV carried out the data collection procedure and formatting the data set. NDM carried out the basic analysis for the study. CR participated in the analysis using software. HBC conceived of the study and coordinating the files.

## Compliance with ethical standards

**Conflict of interest:** Authors do not have any conflict of interest to declare.

**Ethical issues:** None

## References

1. Tamil Nadu Agricultural University. Plant growth and development. Coimbatore: Tamil Nadu Agricultural University; 2023.
2. Ministry of Agriculture & Farmers Welfare. Black gram cultivation practices. New Delhi: Government of India; 2023.
3. Food and Agriculture Organization. Black gram production and consumption trends in India. Rome: FAO; 2023.
4. India Meteorological Department. Impact of weather variability on pulse production in India (IMD Report No. 5678). New Delhi: IMD; 2023.
5. Directorate of Economics and Statistics. Agricultural production

trends in India: 2023-24 outlook. New Delhi: Ministry of Agriculture & Farmers Welfare; 2022.

6. United States Department of Agriculture. Folic acid and maternal health: a nutritional overview. Washington (DC): USDA; 2023.
7. Prioty JK, Rahman KS, Miah MAM. Growth and instability analysis of black gram (*Vigna mungo* L.) in Bangladesh. Bangladesh J Agri. 2023;48(2):30–8. https://doi.org/10.3329/bjagri.v48i2.70156
8. Kumar KM, Satyanarayana PV, Babu PU, Sreenivas G, Manojkumar D. Principal component analysis for yield in black gram genotypes (*Vigna mungo* L. Hepper) under rice fallow pulse cropping system. Biol Forum. 2023;15(3):930–93.
9. Montgomery DC, Peck EA, Vining GG. Introduction to linear regression analysis. 5th ed. Hoboken (NJ): Wiley; 2012.
10. Kutner MH, Nachtsheim CJ, Neter J. Applied linear regression models. 4th ed. New York: McGraw-Hill/Irwin; 2004.
11. Jolliffe IT. Principal component analysis. 2nd ed. New York: Springer; 2002. https://doi.org/10.1007/b98835
12. Abdi H, Williams LJ. Principal component analysis. Wiley Interdiscip Rev Comput Stat. 2010;2(4):433–59. https://doi.org/10.1002/wics.101
13. Mohanlal VA, Saravanan K, Sabesan T. Application of principal component analysis (PCA) for black gram (*Vigna mungo* L.) germplasm evaluation under normal and water-stressed conditions. Legume Res. 2023;46(9):1134–40. https://doi.org/10.18805/LR-4427
14. Kumar P, Patel R, Reddy S. Performance of multiple linear regression and PCA in predicting the yield of legumes. J Agric Stat. 2017;65(3):152–8.
15. Sahu P, Kumar A, Patel R. Evaluation of multicollinearity in crop yield prediction: A comparison of PCA and MLR. Indian J Agric Sci. 2018;88(5):785–92. https://doi.org/10.56093/ijas.v88i5.81629
16. Reddy AK, Priya MS, Reddy DM, Reddy BR. Principal component analysis for yield in black gram (*Vigna mungo* L. Hepper) under organic and inorganic fertilizer managements. Int J Plant Soil Sci. 2021;33(9):26–34. https://doi.org/10.9734/ijpss/2021/v33i930463
17. Reni YP, Ramana MV, Rajesh AP, Madhavi GB, Prakash KK. Principal component analysis for yield and quality traits of black gram (*Vigna mungo* L. Hepper). Int J Plant Soil Sci. 2022;34(7):38–47. https://doi.org/10.9734/ijpss/2022/v34i730887
18. Sharma A, Singh B, Patel P. Application of principal component analysis in crop yield prediction. Int J Agric Sci. 2020;18(4):53–8. https://doi.org/10.15740/HAS/IJAS/18.4/53-58
19. Kumar P, Patel R, Reddy S. Performance of multiple linear regression and PCA in predicting the yield of legumes. J Agric Stat. 2017;65(3):152–8. https://doi.org/10.5958/0976-4666.2017.00021.1
20. Jain A, Kumar S, Singh B. Principal component analysis for crop yield prediction in pulses. Agric Sci Rev. 2021;12(1):44–50. https://doi.org/10.5958/0976-0571.2021.00007.9
21. Sahu P, Kumar A, Patel R. Evaluation of multicollinearity in crop yield prediction: A comparison of PCA and MLR. Indian J Agric Sci. 2018;88(5):785–92. https://doi.org/10.56093/ijas.v88i5.81629
22. Sruthi SR, Laleeth Kumar N, Kishore K, Johnny SI, Anbuselvam Y. Principal component analysis in rice (*Oryza sativa* L.) varieties for three seasons in Annamalai Nagar, an east coast region of Tamil Nadu. Plant Sci Today. 2024. https://doi.org/10.14719/pst.4105

## Additional information

**Peer review**: Publisher thanks Sectional Editor and the other anonymous reviewers for their contribution to the peer review of this work.

**Reprints & permissions information** is available at https://horizonepublishing.com/journals/index.php/PST/open_access_policy

**Publisher's Note**: Horizon e-Publishing Group remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Indexing**: Plant Science Today, published by Horizon e-Publishing Group, is covered by Scopus, Web of Science, BIOSIS Previews, Clarivate Analytics, NAAS, UGC Care, etc
See https://horizonepublishing.com/journals/index.php/PST/indexing_abstracting



**Publisher information:** Plant Science Today is published by HORIZON e-Publishing Group with support from Empirion Publishers Private Limited, Thiruvananthapuram, India.